Florida
Standards Assessments

Florida Standards Assessments

2016–2017

Volume 1
Annual Technical Report

FLORIDA DEPARTMENT OF EDUCATION
fldoe.org

ACKNOWLEDGMENTS

This technical report was produced on behalf of the Florida Department of Education. Requests for additional information concerning this technical report or the associated appendices should be directed to Dr. Salih Binici at the Florida Department of Education (Salih.Binici@fldoe.org).

Major contributors to this technical report include the following staff from American Institutes for Research (AIR): Dr. Harold Doran, Dr. Elizabeth Ayers-Wright, Dr. Dipendra Subedi, Dr. MinJeong Shin, Dr. Ah-Young Shin, Patrick Kozak, Mayumi Rezwan, Kathryn Conway, and Emily Rubenstein. The major contributors from the Florida Department of Education are as follows: Dr. Salih Binici, Dr. Qian Liu, Vince Verges, Susie Lee, Jenny Black, Zhongtian Lin, Susan Taylor, Sally Rhodes, and Travis Barton.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The Florida Standards Assessments (FSA) 2016–2017 Technical Report is provided to document all methods used in test construction, psychometric properties of the tests, summaries of student results, and evidence and support for intended uses and interpretations of the test scores. The technical reports are reported as seven separate, self-contained volumes:

1) *Annual Technical Report*. This volume is updated each year and provides a global overview of the tests administered to students each year.
2) *Test Development*. This volume summarizes the procedures used to construct test forms and provides summaries of the item development process.
3) *Standard Setting*. This volume documents the methods and results of the FSA standard setting process.
4) *Evidence of Reliability and Validity*. This volume provides technical summaries of the test quality and special studies to support the intended uses and interpretations of the test scores.
5) *Summary of Test Administration Procedures*. This volume describes the methods used to administer all forms, security protocols, and modifications or accommodations available.
6) *Score Interpretation Guide*. This volume describes the score types reported and describes the appropriate inferences that can be drawn from each score reported.
7) *Special Studies*. During the course of the year, the Florida Department of Education may request technical studies to investigate issues surrounding the test. This volume is a set of reports provided to the department in support of any requests to further investigate test quality, validity, or other issues as identified.

1.1 PURPOSE AND INTENDED USES OF THE FLORIDA STANDARDS ASSESSMENTS

The primary purpose of Florida's K–12 assessment system is to measure students' achievement of Florida's education standards. Assessment supports instruction and student learning, and the results help Florida's educational leadership and stakeholders determine whether the goals of the education system are being met. Assessments help Florida determine whether it has equipped its students with the knowledge and skills they need to be ready for careers and college-level coursework.

Florida's educational assessments also provide the basis for student, school, and district accountability systems. Assessment results are used to determine school and district grades, which give citizens a standard way to determine the quality and progress of Florida's education system. Assessment results are also used in teacher evaluations to measure how effectively teachers move student learning forward. Florida's assessment and accountability efforts have had a significant positive impact on student achievement over time.

The tests are constructed to meet rigorous technical criteria (Standards for Educational and Psychological Testing [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014]) and to ensure that all students have access to the test content via principles of universal design and appropriate accommodations. Information about the FSA standards and test blueprints can be found in Volume 2, Test Development. Additional evidence of content validity can also be found in Section 4 of Volume 4, Evidence of Reliability and Validity. Evidence about the comparability of online and

paper-and-pencil tests can be found in Section 6 of Volume 4, Evidence of Reliability and Validity. The FSA yields test scores that are useful for understanding whether individual students have a firm grasp of the Florida Standards and also whether students are improving in their performance over time. Additionally, scores can be aggregated to evaluate the performance of subgroups and both individual and aggregated scores can be compared over time in program evaluation methods. The reliability of the test scores can be found in Section 3 of Volume 4, Evidence of Reliability and Validity.

The FSA is a criterion-referenced test that is intended is to measure whether or not students have made progress on the Language Arts Florida Standards (LAFS) and the Mathematics Florida Standards (MAFS). As a comparison, norm-referenced tests compare or rank all test takers to one another. The FSA standards and test blueprints are discussed in Volume 2, Test Development.

Table 1 outlines required uses of the FSA.

Table 1: Required Uses and Citations for the FSA

Assessment	Assessment Citation	Required Use	Required Use Citation
Statewide Assessment Program	s. 1008.22, F.S. Rule 1.09422, F.A.C. Rule 1.0943, F.A.C Rule 1.09432, F.A.C.	Third Grade Retention; Student Progression; Remedial Instruction; Reporting Requirements	s. 1008.25, F.S. Rule 6A-1.094221, F.A.C. Rule 6A-1.094222, F.A.C.
		Middle Grades Promotion	s. 1003.4156, F.S.
		High School Standard Diploma	s. 1003.4282, F.S.
		School Grades	s. 1008.34, F.S. Rule 6A-1.09981, F.A.C.
		School Improvement Rating	s. 1008.341, F.S. Rule 6A-1.099822, F.A.C.
		District Grades	s. 1008.34, F.S.
		Differentiated Accountability	s. 1008.33, F.S. Rule 6A-1.099811, F.A.C.
		Opportunity Scholarship	s. 1002.38, F.S.

1.2 Background and Historical Context of Test

To accompany the development of new Florida educational standards, the FSA was designed to measure students' progress in English Language Arts (ELA), Mathematics, and End-of-Course (EOC) tests. The FSA was first administered to students during spring 2015, replacing the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0) in English Language Arts and Mathematics. It is primarily delivered as an online, fixed-form assessment. In spring 2017, the grade 3 and grade 4 Mathematics assessments were transitioned to online delivery. Paper forms were administered to all students taking grade 3 Reading, and paper accommodated versions were available to students whose Individualized Education Plans (IEPs) or Section 504 Plans indicated such a need.

Within the current Florida statewide assessments program, students in grade 3 must score at Level 2 or higher on the FSA ELA grade 3 assessment in order to be promoted to grade 4. Grade 3 students

who score in Level 1 may still be promoted through one of seven Good Cause Exemptions that are addressed in statute and implemented at the district level. Students must score at Level 3 or above on the Grade 10 ELA and Algebra 1 EOC assessments to meet the assessment graduation requirements set in statute. Students who do not score at Level 3 or higher on these assessments have the opportunity to retake the assessments multiple times; may also use concordant scores on the ACT or SAT to meet the Grade 10 ELA requirement; or may earn a comparative passing score on the Postsecondary Education Readiness Test (PERT) for Algebra 1. Also, students' scores on EOC assessments must count for 30% of a student's final course grade for those courses for which a statewide EOC is administered.

In the rest of this section, the transition to FSA will be highlighted. This brief background should establish the legislative and curricular framework for the technical analyses described in the remaining sections of this volume and other volumes of the technical report.

Developments in 2014

In response to Executive Order 13-276, the state of Florida issued an Invitation to Negotiate in order to solicit proposals for the development and administration of new assessments aligned to the Florida Standards in ELA and Mathematics. After the required competitive bid process, a contract was awarded to the American Institutes for Research (AIR) to develop the new Florida Standards Assessments. The new assessments reflect the expectations of the Florida Standards, in large part by increasing the emphasis on measuring analytical thinking.

During summer 2014, psychometricians and content experts from AIR, the Florida Department of Education, and the Department's Test Development Center met to build forms for spring 2015. Because it was necessary to implement an operational test in the following school year, items from the state of Utah's Student Assessment of Growth and Excellence (SAGE) assessment were used to construct Florida's test forms for the 2014–2015 school year. Assessment experts from FDOE, the Department's Test Development Center, and AIR reviewed each item and its associated statistics to determine alignment to Florida's academic standards and to judge the suitability of the statistical qualities of each item. Only those that were deemed suitable from both perspectives were considered for inclusion on Florida's assessments and for constructing Florida's vertical scale.

It is important to note that, in Florida, post-equating is used each year, so all data used for evaluating student performance on the FSA was derived from the Florida population after the spring 2015 administration.

In addition to the operational test items, field test items were embedded onto test forms administered online in order to build the Florida-specific FSA pool for future use. These items were placed onto test forms using an embedded field test design in the same fixed positions across all test forms within a grade. A very large number of items were field tested as described later in this volume in order to build a substantial bank of items to construct future FSA test forms.

It was also necessary to field test a large pool of text-based Writing prompts that could be used for the future FSA ELA tests. This objective was accomplished via a stand-alone Writing field test that occurred during the winter of 2014–2015. A scientific sample of approximately 25,000 students per grade was selected to participate in this field test, and each student responded to two Writing prompts. Approximately 15 prompts were field tested in each grade. Because only one

prompt is used each year, this field test provided data on a large number of prompts for the state. These prompts have been used beginning in spring 2016.

Developments in 2015

The first operational administration of the FSA occurred in spring 2015. Grades 3 and 4 ELA and Mathematics assessments were administered entirely on paper, and all other grades and subjects were administered primarily online, with the exception of Grades 4–7 text-based writing and a small percentage of students in each grade and subject who required paper-based tests as an accommodation in accordance with an IEP or 504 Plan.

Until new performance standards for this test were in place, statutory requirements called for linking 2015 student performance on Grade 3 ELA, Grade 10 ELA, and Algebra 1 to 2014 student performance on Grade 3 and Grade 10 FCAT 2.0 Reading and NGSSS Algebra 1 EOC, respectively. This linking was required to determine student-level eligibility for promotion (Grade 3 ELA) and graduation (Grade 10 ELA and FSA Algebra 1), which are also statutory requirements. This was accomplished using equipercentile linking for Grade 10 ELA and for Algebra 1. Further legislation enacted in spring 2015 changed the promotion requirement for Grade 3 ELA, instead requiring that students scoring in the bottom quintile be identified for districts to use at their discretion in making promotion and retention decisions for that year only.

Existing legislation also prohibits students from being assessed on a grade-level statewide assessment if enrolled in an EOC in the same subject area. The most significant implication of this legislation was that a significant number of students in Grade 8 participated in the Algebra 1 EOC, but not the FSA Grade 8 Mathematics assessment. This will be discussed in more detail in other volumes of the technical report, especially as it relates to the Grades 3–8 Mathematics vertical scale.

During summer 2015, a new vertical scale for grades 3–10 ELA and grades 3–8 Mathematics was established using statistics from the spring 2015 administration. Standard-setting meetings for grades 3–10 ELA, grades 3–8 Mathematics, and EOC Algebra 1, Algebra 2, and Geometry occurred with educators in August and September 2015. The comprehensive process to set performance standards took into account the feedback from more than 400 educators from across the state as well as from members of the community, businesses, and district-level education leaders. Additionally, the commissioner took into account input from the public, who had the opportunity to submit comments at public workshops and via email, online comment forms, and traditional mail over approximately 12 weeks.

Developments in 2016

During spring 2016, the grade 4 Reading portion of the ELA assessment transitioned to an online delivery. A paper form was made available to students whose IEPs or Section 504 Plans indicated such a need.

Equating procedures were implemented in 2016 to ensure comparability between scores in 2015 and 2016. More information about the method and procedure can be found in Section 6.2.

Developments in 2017

During spring 2017, the grades 3 and 4 Mathematics assessments transitioned to an online delivery. A paper form was made available to students whose IEPs or Section 504 Plans indicated such a need.

1.3 PARTICIPANTS IN THE DEVELOPMENT AND ANALYSIS OF THE FLORIDA STATEWIDE ASSESSMENTS

FDOE manages the Florida statewide assessment program with the assistance of several participants, including multiple offices within FDOE, Florida educators, a Technical Advisory Committee (TAC), and vendors. FDOE fulfills the diverse requirements of implementing Florida's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014).

Florida Department of Education (FDOE)

Office of K-12 Student Assessment. The Office of K-12 Student Assessment oversees all aspects of Florida's statewide assessment program, including coordination with other FDOE offices, Florida public schools, and vendors.

Test Development Center. Funded by FDOE via a grant, the Test Development Center (TDC) works with Florida educators and vendors to develop test specifications and test content and to build test forms.

Florida Educators

Florida educators participate in most aspects of the conceptualization and development of Florida assessments. Educators participate in the development of the academic standards, the clarification of how these standards will be assessed, the test design, and the review of test questions and passages.

Technical Advisory Committee

FDOE convenes a panel once a year (two if technical issues/concerns arise) to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Florida testing. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from multiple Florida school districts.

American Institutes for Research

American Institutes for Research (AIR) was the vendor selected through the state-mandated competitive procurement process. AIR was responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the Florida assessments described in this report. All activities were conducted under the close direction of FDOE staff experts. Beginning in summer 2014, AIR became the primary party responsible for executing psychometric operations for the Florida statewide assessments in English Language Arts and Mathematics.

Human Resources Research Organization

Human Resources Research Organization (HumRRO) has provided program evaluation to a wide variety of federal and state agencies as well as corporate and non-profit organizations and foundations. For the Florida statewide assessments, HumRRO conducts independent checks on the equating and linking activities and reports its findings directly to FDOE. HumRRO also provides consultative services to FDOE on psychometric matters.

Buros Institute of Mental Measurements

Buros Institute of Mental Measurements (Buros) provides professional assistance, expertise, and information to users of commercially published tests. For the 2017 Florida statewide assessments, Buros provided independent operational checks on the equating procedures of the FSA, on-site monitoring of Writing hand-scoring activities, and scanning and editing services provided by AIR. Each year, Buros delivers reports on their observations, which are available upon request.

Caveon Test Security

Caveon Test Security analyzes data for the FSA using Caveon Data Forensics™ to identify highly unusual test results for two primary groups: (1) students with extremely similar test scores and (2) schools with improbable levels of similarity, gains, and/or erasures.

1.4 Available Test Formats and Special Versions

The FSA was administered primarily as an online, fixed-form assessment, making use of several technology-enhanced item types. Students in grade 3 Reading were administered paper forms in spring 2017, and students in the remaining grades and EOCs were provided with access to an accommodated form only if such a need was indicated on their IEP or Section 504 Plan.

Administered test forms contained operational items and embedded field-test (EFT) items in pre-determined slots across each form. Operational items were items used to calculate student scores. The EFT items were non-scored items and were used either to populate the FSA test bank for future operational use or to equate the spring 2017 forms to the spring 2015 baseline FSA scale. While there is only one operational form in grades 3–8 Mathematics and 3–10 Reading, there are multiple test forms in order to vary the EFT items on each form and build a large item bank.

Students in grades 4–10 responded to a single text-based Writing prompt, with grades 4–7 Writing administered on paper and grades 8–10 Writing administered online. Writing and Reading item responses were combined so that the data could be calibrated concurrently and subsequently to form an overall English Language Arts (ELA) score. Scale scores for the separate components were not reported. In this document the term *ELA* is used when referring to the combined Reading and Writing score, and *Reading* is used when referring to only the Reading test form or items.

EOC assessments were administered as online, fixed form assessments to students enrolled in Algebra 1, Algebra 2, and Geometry. These tests had multiple operational forms and also contained EFT items to build future test forms as well as items to equate the spring 2017 forms to the spring 2015 baseline FSA scale.

1.5 STUDENT PARTICIPATION

By statute, all Florida public school students are required to participate in the statewide assessments. Students take the FSA Mathematics, Reading, Writing, or EOC tests in the spring. Retake administrations for EOC assessments occur in the summer, fall, and winter, and grade 10 ELA retake administrations only occur in the fall and spring.

Table 2 shows the number of students who were tested and the number of students who were reported in the spring 2017 FSA by grade and subject area. The participation count by subgroup, including gender, ethnicity, special education, and ELL, is presented in Section 9 of this volume.

Table 2: Number of Students Participating in FSA 2016–2017

Mathematics			ELA		
Grade	Number Tested	Number Reported	Grade	Number Tested	Number Reported
3	229,164	228,745	3	228,701	228,166
4	210,568	210,120	4	212,796	207,703
5	214,572	214,004	5	216,297	211,546
6	198,122	196,833	6	207,523	200,977
7	180,958	179,064	7	205,929	198,891
8	134,922	132,952	8	205,472	198,804
Algebra 1	211,079	208,197	9	209,529	199,934
Algebra 2	123,869	122,638	10	209,482	198,691
Geometry	182,899	180,301			

2. RECENT AND FORTHCOMING CHANGES TO THE TEST

The purpose of this section is to highlight any major issues affecting the test or test administration during the course of the year or to highlight and document any major changes that have occurred to the test or test administration procedures over time.

In spring 2017, grades 3 and 4 Mathematics transitioned to an online delivery. No other substantive changes were made impacting the 2016–2017 school year. In June 2017 House Bill 7069 was passed resulting in several changes test administrations. Beginning in Summer 2017 Algebra 2 will no longer be administered. In addition, some grades and subjects will move from online assessments back to paper assessments beginning in Spring 2019.

3. SUMMARY OF OPERATIONAL PROCEDURES

3.1 ONLINE ADMINISTRATION PROCEDURES

Table 3 shows the schedule for the 2016–2017 FSA administration by testing window.

Table 3: Testing Windows by Subject Area

Assessment	Testing Window
Grades 4–7 paper Writing, Grades 8–10 paper Writing	February 27–March 3, 2017
Grades 8–10 online Writing	February 27–March 10, 2017
Grades 3 paper Reading	March 27–April 7, 2017
Grades 4–10 online Reading, Grades 3–8 online Mathematics	April 10–May 12, 2017
Grades 4–10 paper Reading, Grades 3–8 paper Mathematics	April 10–May 12, 2017
Algebra 1, Algebra 2, and Geometry online	April 17–May 12, 2017
Algebra 1, Algebra 2, and Geometry paper	April 17–May 12, 2017

In accordance with state law, students were required to participate in the spring assessment, and all testing took place during the designated testing window. The FSA tests were administered in sessions, with each session having a time limit. Once a session was started, a student was required to finish it before he or she was permitted to leave the school's campus. A student could not return to a session once he or she left campus.

The key personnel involved with the FSA administration included the district assessment coordinators (DACs), school administrators, and test administrators (TAs) who proctored the test. An online TA training course was available to TAs. More detailed information about the roles and responsibilities of various testing staff can be found in Volume 5 of the 2017 FSA Annual Technical Report.

A secure browser developed by AIR was required to access the online FSA tests. The browser provided a secure environment for student testing by disabling the hot keys, copy, and screenshot capabilities and blocking access to desktop functionalities, such as the Internet and e-mail. Other measures that protected the integrity and security of the online test are presented in Volume 5 of the 2017 FSA Technical Report.

3.2 ACCOMMODATIONS FOR FSA

Florida assessments are inclusive for all students, which serves as evidence of test validity. To maximize the accessibility of the assessments, various accommodations were provided to students with special needs, as indicated by documentation such as IEPs or Section 504 Plans. Such accommodations improve the access to state assessments and help students with special needs demonstrate what they know and are able to do. From the psychometric point of view, the purpose of providing accommodations is to "increase the validity of inferences about students with special needs by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

The paper version is constructed to the exact same test specifications and, in many cases, the items between the online and paper forms are exactly the same. Some technology enhanced items are replaced on the paper versions with items intended to render on paper. They are chosen to essentially mirror the online items they are replacing such that the paper form measures the same construct in a similar way.

Observed data collected from the test administrations provide evidence that the test forms are equally as reliable and that students on the paper form also have a range of scores. This evidence indicates that high performing students taking an accommodated form can still demonstrate high performance and they are not impeded in any way by the nature of the form or its administration. Cronbach's alpha by reporting category are given for online and accommodated groups in Tables 11 – 50 of Appendix A of Volume 4.

The number of students who took the paper-and-pencil version of the 2016–2017 FSA varies between 263 and 1,626 cross grades and subjects, as shown in Table 4.

Table 4: Counts of Paper-and-Pencil Assessments by Grades and Subjects

Subject	Grade	Spring 2017
Mathematics	3	1,626
	4	1,316
	5	1,537
	6	758
	7	687
	8	693
EOC	Algebra 1	679
	Algebra 2	263
	Geometry	650
Reading	4	1,327
	5	1,534
	6	765
	7	710
	8	766
	9	808
	10	924

Table 5 shows the percent of students in each performance level for grades and subjects that had paper accommodated forms in Spring 2017.

Table 5: Percent of Students Taking Paper Forms by Performance Level

Subject	Grade	Level 1	Level 2	Level 3	Level 4	Level 5
Mathematics	3	42.1	20.7	23.0	11.7	2.4
	4	46.5	20.6	19.6	10.8	2.6
	5	50.0	23.0	16.4	8.0	2.6
	6	54.8	22.3	15.0	5.9	2.0
	7	49.9	22.8	17.8	8.2	1.3
	8	52.1	23.1	16.9	4.9	3.0
EOC	Algebra 1	62.6	12.5	16.1	6.2	2.7
	Algebra 2	49.2	17.9	21.4	5.0	6.5
	Geometry	57.5	17.1	18.3	4.4	2.7
ELA	4	46.2	27.5	17.8	7.1	1.4
	5	52.2	28.5	13.8	4.5	0.9
	6	54.5	25.0	12.7	6.3	1.5
	7	53.3	22.7	13.8	8.0	2.2
	8	47.4	21.3	18.2	9.3	3.6
	9	56.1	20.4	12.0	8.0	3.6
	10	51.2	25.2	11.4	9.8	2.4

The test administrator and the school assessment coordinator were responsible for ensuring that arrangements for accommodations were made before the test administration dates. For eligible students participating in paper-based assessments, a variety of accommodations were available, such as large print, contracted braille, uncontracted braille, and displaying only one item per page. For eligible students participating in computer-based assessments, accommodations such as masking, text-to-speech, and regular or large-print passage booklets were made available. Students had the opportunity to use these accommodations only as dictated on their IEPs or Section 504 Plans. Additional accommodations and further explanation of the guidelines can be found in Volume 5, Summary of Test Administration Procedures.

4. MAINTENANCE OF THE ITEM BANK

4.1 OVERVIEW OF ITEM DEVELOPMENT

Complete details of AIR's item development plan are provided in the 2016–2017 Annual Technical Report, Volume 2, Test Development. The test development phase included a variety of activities designed to produce high-quality assessments that accurately measure skills and abilities of students with respect to the academic standards and blueprints.

New items are developed each year to be added to the operational item pool after being field tested. Several factors determine the development of new items. The item development team conducts a gap analysis for distributions of items across multiple dimensions, such as item counts, item types, item difficulty, depth of knowledge (DOK) levels, and numbers in each strand or benchmark.

In spring 2017, field-test items were embedded on online forms. Future FSA items were not field tested on paper this year, so there were no field-test items in grade 3 Reading. All assessments were fixed-form with a predetermined number and location of field-test items. The paper accommodated versions of online assessments contained filler items in the field-test slots to ensure equal length assessments. These items were not analyzed as part of field-test calibrations.

4.2 REVIEW OF OPERATIONAL ITEMS

During operational calibration, items were reviewed based on their performance during the spring administration. In some instances, operational items were removed from scoring based on content or statistical anomalies that were not apparent during form building.

Prior to the spring administration, a *Calibration and Scoring Specifications* document was created by AIR, FDOE, and HumRRO and reviewed by the TAC. The specifications document outlined all details of item calibration, flagging rules for items, equating to the Spring 2015 baseline scale, pre-equating of paper accommodated forms, and scoring. AIR used the specifications to complete classical item analyses and IRT calibrations (see Sections 5 and 6 of this volume) for each test and posted results to a secure location for review. During the spring calibrations, daily calls were scheduled that included all parties: AIR, FDOE, TDC, HumRRO, and Buros. Items were reviewed, with special attention being paid to items flagged based on the statistical rules described in the Calibration and Scoring Specifications document. These flagging rules are outlined in the sections that follow. Psychometricians and content experts worked together to review items and their statistics to determine if any items were to be removed from scoring.

4.3 FIELD TESTING

The FSA item pool grows each year through the field testing of new items. Any item used on an assessment is field tested before it is used as an operational item. There are two primary ways to field test items: through either an independent field test (IFT) or an embedded field test (EFT).

During the 2016–2017 school year, there were no IFTs.

4.3.1 Embedded Field Test

FSA forms were pre-built with approximately 10 field-test items embedded onto each test form, and each form was assigned to students randomly as described below. Some field-test items appeared on multiple forms.

Table 6 shows the number of Mathematics and EOC items by grade and item type that were included on forms for field testing. Table 7 shows the number of Reading items by grade and item type that were included on forms for field testing. During calibrations, some items were dropped from the initial item pool due to poor performance. Appendix B provides the number of field test items remaining after removal of items during calibrations.

Table 6: Mathematics and EOC Field-Test Items by Item Type and Grade

Item Type	3	4	5	6	7	8	Algebra 1	Algebra 2	Geometry
MC4	73	46	18	22	26	39	45	35	23
MS5	29	37	5	5	3	11	5	6	4
MS6	4	17	1	3	4	4	3	3	1
GRID	39	25	3	16	11	17	4	8	5
Hot Text	0	0	0	0	0	2	6	5	4
EQ	135	141	36	39	82	65	19	38	39
NL	1	0	0	0	2	0	0	0	0
Editing Task Choice	0	0	0	0	0	0	6	4	1
Match	9	22	7	2	4	2	0	1	0
Table	15	8	1	4	3	4	2	0	0

Table 7: Reading Field Test Items by Item Type and Grade

Item Type	4	5	6	7	8	9	10
MC4	45	20	19	33	21	21	38
MS5	0	1	2	6	1	3	5
MS6	2	1	0	0	0	0	0
Editing Task Choice	14	8	0	4	6	0	0
Hot Text	8	3	2	7	6	6	2
EBSR4	13	4	11	9	8	8	15
EBSR5	1	1	0	0	0	1	1
EBSR6	1	1	0	1	0	0	0

With fixed-form assessments, it is known how many items are unique to a form. Thus, based on the number of students participating, as well as the number of forms, the expected number of responses per item can be calculated.

The form distribution algorithm employed by AIR ensures that forms are drawn and assigned to students according to a simple random sample. For example, suppose there are *J* total forms in the pool, items appear on only one form, and a total of *N* students are participating in the field test. The probability that any one of the *J* forms can be assigned to one student is *1/J*. Thus, the expected number of student responses for each form is

$$S = \frac{N}{J},$$

where *J* is the number of forms in the pool, *N* is the number of students who will be participating in the field test, and *S* is the sample size per item. If an item appears on more than one form, the expected sample size would be *S* times the number of forms on which the item appears.

The aim was to achieve a minimum sample size of 1,500 students per item. Hence, given a test length of *L* and fixing *S* at 1500 (the expected sample size per item), we can determine the maximum number of forms that can exist in the pool as

$$J = \frac{N}{1500}.$$

From this, we see that

- a random sample of students receives each form; and
- for any given form, the students are sampled with equal probability.

Table 8 and Table 9 show the total number of forms administered in spring 2017. In each grade, there was a single core or operational form. The same core form was replicated for each anchor or embedded field-test form, resulting in multiple forms for each grade and subject. For the EOCs, there were multiple core forms, each also replicated to create a number of embedded field-test forms.

Table 8: ELA Form Summary

Grade	Total Number of Forms
3	3
4	16
5	8
6	8
7	12
8	8
9	9
10	13

Table 9: Mathematics and EOC Form Summary

Grade	Total Number of Forms
3	39
4	35
5	14
6	14
7	18
8	19
Algebra 1	11
Algebra 2	10
Geometry	8

A detailed overview of the development and review process for new items is given in the 2016–2017 FSA Technical Report, Volume 2, Test Development. Additional details on development and maintenance of the item pool are also given in the same volume.

5. ITEM ANALYSES OVERVIEW

5.1 CLASSICAL ITEM ANALYSES

Item analyses examine whether test items function as intended. Overall, a minimum sample of 1,500 responses (Kolen & Brennan, 2004) per item was required for both classical analysis and for the item response theory (IRT) analysis. However, many more responses than 1,500 were always available. For operational item calibrations, an early processing sample was used in the analyses; for field-test item calibrations, all students were used. Similarly, a minimum sample of 200 responses (Zwick, 2012) per item in each subgroup was applied for differential item functioning (DIF) analyses.

Several item statistics were used to evaluate multiple-choice (MC) and non-multiple choice items, generally referred to as constructed response (CR), for integrity and appropriateness of the statistical characteristics of the items. The thresholds used to flag an item for further review based on classical item statistics are presented in Table 10.

Table 10: Thresholds for Flagging Items in Classical Item Analysis

Analysis Type	Flagging Criteria
Item Discrimination	Point biserial correlation for the correct response is < 0.25
Distractor Analysis	Point biserial correlation for any distractor response is > 0
Item Difficulty (MC items)	The proportion of students (*p*-value) is < 0.20 or > 0.90
Item Difficulty (non-MC items)	Relative mean is <0.15 or >0.95

Item Discrimination

The item discrimination index indicates the extent to which each item differentiated between those examinees who possessed the skills being measured and those who did not. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. The discrimination index for multiple-choice items was calculated as the correlation between the item score and the ability estimate for students. Point biserial correlations for operational items can be found in Appendix A.

Distractor Analysis

Distractor analysis for multiple-choice items was used to identify items that may have had marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracted high-scoring students. For multiple-choice items, the correct response should have been the most frequently selected option by high-scoring students. The discrimination value of the correct response should have been substantial and positive, and the discrimination values for distractors should have been lower and, generally, negative.

Item Difficulty

Items that were either extremely difficult or extremely easy were flagged for review but were not necessarily deleted if they were grade-level appropriate and aligned with the test specifications. For multiple-choice items, the proportion of students in the sample selecting the correct answer

(the *p*-value) was computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty was calculated using the item's relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the maximum possible score points). Conventional item *p*-values and IRT parameters are summarized in Section 6.4. The *p*-values for operational items can be found in Appendix A.

5.2 DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) document provides a guideline for when sample sizes permitting subgroup differences in performance should be examined and when appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors. To identify such potential problems, FSA items were evaluated in terms of DIF statistics.

DIF analysis was conducted for all items to detect potential item bias across major gender, ethnic, and special population groups. Because of the limited number of students in some groups, DIF analyses were performed for the following groups:

- Male/Female
- White/African-American
- White/Hispanic
- Student with Disability (SWD)/Not SWD
- English Language Learner (ELL)/Not ELL

Differential item functioning (DIF) refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF was important because it provided a statistical indicator that an item may contain cultural or other bias. DIF-flagged items were further examined by content experts who were asked to reexamine each flagged item to make a decision about whether the item should have been excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF. For example, if schools in certain areas are less likely to offer rigorous Geometry classes, students at those schools might perform more poorly on Geometry items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but rather the instruction. However, DIF can indicate bias, so all items were evaluated for DIF.

A generalized Mantel-Haenszel (MH) procedure was applied to calculate DIF. The generalizations include (1) adaptation to polytomous items and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's raw score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the MHχ^2 DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the $MH\chi^2$ value, the

conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

The MH chi-square statistic (Holland and Thayer, 1988) is calculated as

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})}$$

where $k = \{1, 2, \ldots K\}$ for the strata, n_{R1k} is the number of correct responses for the reference group in stratum k, and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k} n_{R+k}}{n_{++k}}$$

where n_{+1k} is the total number of correct responses, n_{R+k} is the number of students in the reference group, and n_{++k} is the number of students, in stratum k, and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k} n_{F+k} n_{+1k} n_{+0k}}{n_{++k}^2 (n_{++k} - 1)}$$

n_{F+k} is the number of students in the focal group, n_{+1k} is the number of students with correct responses, and n_{+0k} is the number of students with incorrect responses, in stratum k.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k} n_{F0k} / n_{++k}}{\sum_k n_{R0k} n_{F1k} / n_{++k}}.$$

The MH-delta (Δ_{MH}, Holland & Thayer, 1988) is then defined as

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH}).$$

The GMH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)' \left(\sum_k var(\boldsymbol{a}_k)\right)^{-1} \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)$$

where $\boldsymbol{a}_k$ is a $(T-1)\ X\ 1$ vector of item response scores, corresponding to the T response categories of a polytomous item (excluding one response). $E(\boldsymbol{a}_k)$ and $var(\boldsymbol{a}_k)$, a $(T-1) \times (T-1)$ variance matrix, are calculated analogously to the corresponding elements in $MH\chi^2$, in stratum k.

The standardized mean difference (SMD, Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK} m_{FK} - \sum_k p_{FK} m_{RK}$$

where

$$p_{FK} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum k,

$$m_{FK} = \frac{1}{n_{F+k}}\left(\sum_t a_t n_{Ftk}\right)$$

is the mean item score for the focal group in stratum k, and

$$m_{RK} = \frac{1}{n_{R+k}}\left(\sum_t a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum k.

Items were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. DIF classification rules are illustrated in Table 11. Items were also indicated as positive DIF (i.e., +A, +B, or +C), signifying that the item favored the focal group (e.g., African-American, Hispanic, or female) or negative DIF (i.e., –A, –B, or –C), signifying that the item favored the reference group (e.g., white or male). If the DIF statistics fell into the "C" category for any group, the item showed significant DIF and was reviewed for potential content bias or differential validity, whether the DIF statistic favored the focal or the reference group. Content experts reviewed all items flagged on the basis of DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance in field testing.

Table 11: DIF Classification Rules

Dichotomous Items	
Category	*Rule*
C	MH_{X^2} is significant and $\lvert\hat{\Delta}_{MH}\rvert \geq 1.5$
B	MH_{X^2} is significant and $1 \leq \lvert\hat{\Delta}_{MH}\rvert < 1.5$
A	MH_{X^2} is not significant or $\lvert\hat{\Delta}_{MH}\rvert < 1$
Polytomous Items	
Category	*Rule*
C	MH_{X^2} is significant and $\lvert SMD\rvert / \lvert SD\rvert > .25$
B	MH_{X^2} is significant and $.17 < \lvert SMD\rvert / \lvert SD\rvert \leq .25$
A	MH_{X^2} is not significant or $\lvert SMD\rvert / \lvert SD\rvert \leq .17$

DIF summary tables can be found in Appendix A for operational items and in Appendix B for field-test items. Across all operational items and DIF comparison groups, less than 1% of Mathematics and EOC items were classified as C DIF and less than 1% of ELA items were classified as C DIF. Items were reviewed by content specialists and psychometricians to ensure that they were free of bias.

Across all field-test items and DIF comparison groups, less than 2% of Mathematics and EOC items were classified as C DIF, and 1% of ELA items were classified as C DIF. All field-test items will be reviewed by content specialists and psychometricians before being placed on forms for operational use. More information about test construction and item review can be found in Volume 2.

In addition to the classical item summaries described in this section, two IRT-based statistics were used during item review. These methods are described in Section 6.3.

6. ITEM CALIBRATION AND SCALING

Item Response Theory (IRT; van der Linden & Hambleton, 1997) was used to calibrate all items and derive scores for all FSA tests. IRT is a general framework that models test responses resulting from an interaction between students and test items. One advantage of IRT models is that they allow for item difficulty to be scaled on the same metric as person ability.

IRT encompasses a large number of related measurement models. Models can be grouped into two families. While both families include models for dichotomous and polytomous items, they differ in their assumptions about how student ability interacts with items. The Rasch family of models includes the Rasch model and the Master's Partial Credit Model. The Rasch family is distinguished in that models do not incorporate a pseudo-guessing parameter and it assumes that all items have the same discrimination.

Extensions to the Rasch model include the 2- and 3-parameter logistic (2PL, 3PL) models and the Generalized Partial Credit Model. These models differ from the Rasch family of models by including a parameter that accounts for the varied slopes between items, and in some instances, models also include a lower asymptote that varies to account for pseudo-guessing that may occur with some items. A discrimination parameter is included in all models in this family and accounts for differences in the amount of information items may provide along different points of the ability scale (the varied slopes). The 3PL is characterized by a lower asymptote, often referred to as a *pseudo-guessing parameter*, which represents the minimum expected probability of answering an item correctly. The 3PL is often used with multiple-choice items, but it can be used with any item where there is a possibility of guessing.

Operational item calibrations were completed on an Early Processing Sample (EPS) collected during the spring administration. The EPS was a representative, scientific sample of students across the state. The sampling of students was accomplished using a stratified random sample with explicit and implicit strata that were chosen to represent important characteristics of the tested student population. Region was used as explicit strata, whereas gender, ethnicity, school size, mean theta score, and curriculum group (Standard, LEP, ESE) were used as implicit strata. The *region* variable is intended to capture the differences in student population across the state. Male and female are the subgroups under *gender*, whereas *ethnicity* is comprised of white, African American, Hispanic, and others subgroups. *Mean theta score* provides the measure of the student ability across the population based on the previous year's data. The *school size* variable is used in sampling to ensure that the sample is composed of schools of various sizes. The *curriculum group* variable has three subgroups: Standard, Limited English Proficiency (LEP), and Exceptional Student Education (ESE). This variable shows that the representativeness of ELL population is also evaluated as part of the sample evaluation. More information about the EPS can be found in Appendix C.

FDOE and AIR collaborated through several rounds of review to ensure that the strata were appropriately defined and the student population was adequately represented; this EPS plan, which can be found in Appendix C, was also reviewed and affirmed by the TAC. For grade 8 Mathematics and EOC calibrations, the entire population was used instead of the EPS.

Two general approaches are used in IRT to calibrate items and score students based on the estimated item difficulties. In pre-equating, item responses are collected from a student group, the

statistical characteristics of the items are estimated from that group, and then these statistics are used to score all future groups of students. This approach assumes that the characteristics of the items remain constant over time. A second approach is post-equating. In this approach, item responses are collected from a student group, and the statistical characteristics of the items are re-estimated from those responses. However, these statistical characteristics are assumed to apply only to this student group. New item statistics are collected each year when items are used, thus assuming the statistical characteristics of the item may change when the tested students change

In Florida, this second approach of post-equating was used, and all data regarding item responses were derived from the most recent group of students to be administered the test. Beginning in 2016, test forms were equated to the Spring 2015 FSA scale, a step that was not necessary in the initial year. This process is described in further detail in Section 6.2.

Field-test item calibrations were completed on the entire sample from the spring administration to ensure adequate sample sizes for all items. Field-test items were equated to the operational scale using the Stocking-Lord procedure.

6.1 ITEM RESPONSE THEORY METHODS

The generalized approach to item calibration was to use the 3-parameter logistic model (3PL; Lord & Novick, 1968) for multiple-choice items; to use the 2-parameter logistic model (2PL; Lord & Novick, 1968) for binary items that assume no guessing; and to use the Generalized Partial Credit Model (GPCM; Muraki, 1992) for items scored in multiple categories.

For items with some probability of guessing, such as multiple-choice items, the 3PL model was used, since it incorporates a parameter to account for guessing. For non-multiple-choice binary items, the content of the item was reviewed. If it was determined that there was no probability of guessing, then the 2PL model was used; however, the 3PL model was used if guessing was in fact possible.

The 3-parameter model is typically expressed as

$$\mathrm{P_i}(\theta_j) = \mathrm{c_i} + \frac{1 - \mathrm{c_i}}{1 + \exp\left[-\mathrm{Da_i}\left(\theta_j - \mathrm{b_i}\right)\right]}$$

where $\mathrm{P_i}(\theta_j)$ is the probability of examinee j answering item i correct($\mathrm{c_i}$)is the lower asymptote of the item response curve (the pseudo-guessing parameter) $\mathrm{b_i}$ is the location parameter, $\mathrm{a_i}$ is the slope parameter (the discrimination parameter), and D is a constant fixed at 1.7 bringing the logistic into coincidence with the probit model. Student ability is represented by θ_j. For the 2PL, the pseudo-guessing parameter ($\mathrm{c_i}$) is set to 0.

The Generalized Partial Credit Model is typically expressed as the probability for individual j of scoring in the $(z_i + 1)$th category to the ith item as

$$\mathrm{P}\left(z_i \middle| \theta_j\right) = \frac{\exp \sum_{k=0}^{z_i} \mathrm{Da_i}\left(\theta_j - \delta_{ki}\right)}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} \mathrm{Da_i}\left(\theta_j - \delta_{ki}\right)}$$

where δ_{ki} is the kth step value, $z_i = 0,1,..,m_i$, m_i is the maximum possible score of the item and $\sum_{k=0}^{0} \mathrm{Da_i}\left(\theta_j - \delta_{ki}\right) \equiv 0$.

All item parameter estimates were obtained with IRTPRO version 4.0 (Cai, Thissen, & du Toit, 2011). IRTPRO uses marginal maximum likelihood estimation (MLE).

6.2 EQUATING TO THE 2015 SCALE

Post-equating the item parameters to the 2015 baseline scale was a new task in 2016 and was also performed in 2017. Equating is a procedure in which test scores from different test instruments are placed onto a common scale so that scores from different test administrations can be directly compared. Equating in Florida begins with the development of an "anchor" set - a set of items that are common between two different versions of a test. This is the basis of the common-item non-equivalent groups anchor design (Kolen & Brennan, 2004). This anchor set is essentially a miniature version of a parallel test with respect to its content and statistical characteristics. That is, the items in the anchor set represent the blueprint percentages as well as having similar statistical properties as the assessment the test is being linked to.

During test construction, items are selected and evaluated using their statistical properties collected from the test bank. These statistical characteristics are provisional, given that post-equating is used, but they are useful for guiding the construction of anchor sets, as well as the overall test form. The statistical characteristics typically include evaluations of an item's *p*-value, point-biserial correlations, and IRT-based characteristics (i.e., difficulty, guessing, slope), and differential item functioning (DIF). Items are selected such that forms meet the same blueprint as the baseline year, and classical and IRT summary statistics are also calculated and compared to the baseline year. The process is iterative and continues to choose items with content and statistical properties, as well as professional judgment by content experts, to build a linking set that conforms to the blueprint and statistical characteristics of the baseline year forms. Once finalized, a subset of items are labeled as *anchor* items to be used to complete equating during operational calibrations. Additional details about test construction are available in Volume 2, Test Construction.

6.2.1 Online Forms

Online operational and anchor items were jointly analyzed using the early processing sample (EPS) in Mathematics grades 3-7 and ELA grades 3-10 and using the entire population in Mathematics grade 8 and in the EOCs. The EPS is a scientific sample of students and is representative of the students in the state of Florida. Prior to analyses, demographics of the EPS were compared to state values used to draw the samples to ensure representativeness. More information about the EPS can be found in Appendix C. HumRRO replicated all item calibrations and provided an independent list of flagged items. Buros provided additional commentary on calibrations and flagged items.

Classical item statistics, as described in Section 5, were computed first and reviewed to determine if any items should be removed from analyses prior to either IRT calibrations or equating. Content experts from both AIR and TDC reviewed flagged items to ensure that they were being scored correctly. IRT calibrations were then performed, and item summaries, as described in Section 6.3, were calculated. Items with anomalous parameters or flagged for item fit were reviewed by psychometricians and content experts. Any item found to be misbehaving was dropped, and the IRT calibration was then rerun. A sample calibration report and sample anchor item summary report produced by AIR can be found in Appendix G. Once a final IRT calibration was determined, all parties could proceed with the initial equating solution.

Using the calibrated item statistics from IRTPRO, the complete set of equating items (all internal and external anchor items) was used to calculate the equating constants to place the 2017 item parameters onto the 2015 scale. Internal anchor items are operational and are used to calculate student scores. External anchor items are located in embedded field test slots and do not count toward student scores. The Stocking-Lord procedure was used to complete the equating.

The Stocking-Lord (Stocking & Lord, 1983) procedure is a method commonly used alongside the 3-parameter logistic model and Generalized Partial Credit Model and establishes the linking constants, *A* and *B*, that minimize the squared distance between two test characteristic curves. *A* is often referred to as the *slope* and *B* is often referred to as the *intercept*. The symmetric approach evaluates the following integral, where the index *i* denotes a common item, and subscripts *I* and *J* denote the item parameters for the bank and item parameters to be rescaled:

$$\arg\min SL = \int \left[\sum_{i=1}^{K} E\left(z_{i,I}|\theta_1\right) - \sum_{i=1}^{K} E\left(z^*_{i,J}|\theta_1\right)\right]^2 f(\theta_1|\mu,\sigma^2)\, d\theta_1 + \int \left[\sum_{i=1}^{K} E\left(z^*_{i,I}|\theta_2\right) - \sum_{i=1}^{K} E\left(z_{i,J}|\theta_2\right)\right]^2 f(\theta_2|\mu,\sigma^2)\, d\theta_2$$

where $f(\theta_1|\mu,\sigma^2)$ is the normal population density associated with putting operational items onto the bank scale and $f(\theta_2|\mu,\sigma^2)$ is the density associated with putting bank items onto the operational scale. Without loss of generality to permit for compact notation, let $E\left(z_{i,I}|\theta\right)$ denote the expected value of response on the ith item from either the binary or partial credit model and let $E\left(z_{i,J}|\theta\right)$ be the same for the items to be rescaled.

Where for dichotomous items we have

$$p\left(z_{i,I}=1|\theta\right) = \mathrm{c}_{\mathrm{i},I} + \frac{1-\mathrm{c}_{\mathrm{i},I}}{1+\exp[-\mathrm{Da}_{\mathrm{i},I}\left(\theta-\mathrm{b}_{\mathrm{i},I}\right)]}$$

and for the polytomous IRT models

$$p\left(z_{i,I}|\theta\right) = \frac{\exp\left(\sum_{\mathrm{k}=0}^{\mathrm{z_i}} \mathrm{Da_i}\left(\theta-\delta_{\mathrm{ki},I}\right)\right)}{\sum_{\mathrm{h}=0}^{\mathrm{m_i}} \exp\sum_{\mathrm{k}=0}^{\mathrm{h}} \mathrm{Da}_{\mathrm{i},I}\left(\theta-\delta_{\mathrm{ki},I}\right)}$$

where z_i denotes score point $\mathrm{z_i}=\{1,\ldots,\mathrm{m}_i\}$ to item i. The expected score for the polytomous models is

$$E\left(z_{i,I}|\theta\right) = \sum_{z=1}^{m_i} zp(\mathrm{z}_i|\theta).$$

The symmetric approach uses the reverse transform for the bank items

$$p\left(z^*_{i,I}=1|\theta\right) = \mathrm{c}_{\mathrm{i},I} + \frac{1-\mathrm{c}_{\mathrm{i},I}}{1+\exp[-\mathrm{DAa}_{\mathrm{i},I}\left(\theta-\frac{(\mathrm{b}_{\mathrm{i},I}-B)}{A}\right)]}$$

and for the polytomous IRT models

$$p\left(z_{i,I}^{*}|\theta\right) = \frac{exp\left(\sum_{k=0}^{z_i} DAa_i\left(\theta - \frac{(\delta_{ki,I} - B)}{A}\right)\right)}{\sum_{h=0}^{m_i} exp\sum_{k=0}^{h} DAa_{i,I}\left(\theta - \frac{(\delta_{ki,I} - B)}{A}\right)}$$

Once the equating constants were estimated, they were applied to *maximum a posteriori* (MAP) ability estimates, which were derived using IRTPRO, to project the percentage of students, based on the calibration sample, who were likely to score in each performance category. This initial equating solution was referred to as the baseline solution for each grade and subject combination.

After the baseline solution was estimated, two iterative procedures were implemented. The first procedure dropped one item from the equating set per iteration, resulting in a new slope and intercept that was plotted for review. The second procedure started with the baseline solution and cumulatively dropped extreme items from the equating set. This second procedure was implemented via the following steps:

1. Rescaled the 2017 item parameters to be on the 2015 scale using transformation constants based on the Stocking-Lord procedure
2. Computed the weighted area between the item characteristic curves (ICC), a method known as D^2 or the mean squared difference (MSD)
3. Computed the mean and standard deviation of the MSD and standardized the MSD to get SMSD (standardized MSD)
4. Ordered all equating items by |SMSD|
5. Identified "extreme" items as any item with |SMSD| > 2.5 and removed item(s) with max|SMSD| from equating set
6. Iteratively removed items in the linking set until $SMSD < 2.5 \; \forall \; i$

The D^2, or the MSD, is computed by integrating out θ as follows:

$$D^2 = \int \left(E(z_{i,J}|\theta) - E(z_{i,I}|\theta)\right)^2 f(\theta;\mu,\sigma^2)d\theta.$$

The D^2 integral does not have a closed form solution, and so its approximation is based on the weighted summation over q={1, 2, …, 30} quadrature points, all taken from equally spaced points interior to the normal density, *w*, between –4 and 4 of the marginal distribution

$$D^2 = \sum_{q=1}^{30} w_q \left(E\left((z_{i,J}|\theta_q\right) - E\left(z_{i,I}|\theta_q\right)\right)^2.$$

The iterative nature of this process provided for a baseline solution with a projection of its impact on the population percentage of students scoring in each performance level, as well as additional solutions based on the number of extreme items removed. The number of additional solutions conducted depended on the number of extreme items in the equating set. For each additional solution, population impact statistics were provided.

The process described above was automated via AIR's equating software. After the initial equating solution and two iterative procedures were complete, AIR produced an equating report to deliver to FDOE. A sample report can be found in Appendix G. Upon review of these solutions and

discussion during a calibration and equating call including HumRRO and Buros, FDOE and TDC were also able to require removal of additional anchor items based on a variety of factors. These factors included, but were not limited to:

- Content review of any flagged item
- Change in the classical item statistics (e.g., *p*-value)
- Item position shift
- Item fit plots
- Results of anchor item stability checks (e.g., D^2)
- Individual and cumulative impact of anchor items on the scale transformation coefficients
- Evaluation of pre- vs. post-equated item characteristic curves
- Percent of students classified in each achievement level

AIR accommodated requests by dropping items identified by FDOE and TDC following calibration and equating calls. The equating report was updated with new solutions after each request was completed.

Table 12 shows the final equating results. The number of items in the equating design is shown, as well as the number of dropped items and the number of items in the final equating solution. The last two columns show the slope and intercept from the final Stocking-Lord equating solution.

In equating, there are two possible sources of error: sampling error and equating error (Phillips, 2010). Sampling error exists given that calibrations and equating methods are performed on a sample of students drawn from the population. Our sampling design minimizes the design effect that arises from the clustering of students within a group (Phillips, 2010) and uses a stratified random sample of students from across the state. This sampling is described above and in Appendix C.

A second, and potentially larger, source of variance is due to the sampling of common items. The items chosen to link the test forms are only a sample of the items that could have been used to establish the linkage. That is, these items are not treated as fixed, but as a random draw from the universe of potential linking items. Had different items been chosen, a different equating solution would have been found and the degree to which this varies due to the common items can be a very large source of potential error variance (Michaelides and Haertel, 2004). The source of such error is explored during the equating work by dropping items one-by-one from the anchor set and recalculating the slope and intercept. The final distribution of slopes and intercepts was reviewed to see if any single item had a large impact. This process was included in the reports and can be seen on page 29 of Appendix G. Cohen, Johnson, and Angeles (2000) found that the error associated with the uncertainty of IRT parameter estimates resulted in a 25% to 100% increase in standard errors. Error due to the sampling of items is reduced as the number of linking items increases (Michaelides and Haertel, 2004). The uncertainty due to the sampling of items is unaffected by an increase in sample size.

The equating results in the table below represent the final solutions used for FSA equating. The intercept and slope represent the first and second moments of the ability distribution, respectively.

Hence, slope values greater than 1 indicate greater heterogeneity in the population relative to the baseline year, and values less than 1 indicate greater homogeneity than previously observed. Similarly, intercept values greater than 0 indicate an improvement in mean performance relative to the baseline group and values less than 0 denote the opposite.

Table 12: Final Equating Results

Subject	Grade	Number of Items in Design	Number of Items Dropped	Number of Items in Final Solution	Slope	Intercept
ELA	3	33	0	33	0.97493	0.10862
	4	37	0	37	0.96784	0.09542
	5	35	0	35	1.06391	0.00240
	6	34	0	34	1.02280	0.04339
	7	36	0	36	1.01689	0.01826
	8	33	0	33	1.08116	-0.02128
	9	35	0	35	1.04377	0.03307
	10	39	1	38	1.07366	-0.00496
Mathematics	3	40	6	34	0.99335	0.07032
	4	40	1	39	1.00969	0.10867
	5	40	0	40	1.07975	0.07733
	6	40	0	40	1.06391	-0.08132
	7	40	0	40	1.08483	0.03184
	8	40	0	40	1.06339	-0.09673
EOC	Algebra 1	31	0	31	1.03645	-0.11598
	Algebra 2	31	0	31	0.95313	0.38151
	Geometry	31	1	30	1.00153	-0.03120

6.2.2 Paper Accommodated Forms

During Spring 2017, the paper accommodated forms were scored using item parameters from the item's latest online administration. In most instances, parameters from the Spring 2017 online calibrations were applied. In Grades 4-10 ELA all core items are common between online and paper accommodated forms. Thus, all parameters used for scoring were from the Spring 2017 ELA online calibration, including the three Writing prompt items. In Grades 5-8 Mathematics, Algebra 2, and Geometry paper accommodated items that were common with the online form used item parameters from the Spring 2017 online calibrations, and all other items used item parameters from previous online administrations.

In some instances in Grades 3 and 4 Mathematics and Algebra 1, there were paper accommodated items that were not previously calibrated and did not appear on any online form in 2017. These items were placed in field test positions on an online form and were calibrated along with the core items and anchor items. The resulting parameters were then applied to the paper accommodated form.

The paper accommodated forms were automatically equated to the 2015 baseline scale since item parameters came from previously equated Spring 2017, Spring 2016, or Spring 2015 item parameters.

6.3 IRT ITEM SUMMARIES

6.3.1 Item Fit

Yen's Q1 (1981) is used to evaluate the degree to which the observed data fit the item response model. Q1 is a fit statistic that compares observed and expected item performance. To calculate fit statistics before scores were available from AIR's scoring engine, MAP estimates from IRTPRO were used for student ability estimates in the calculations. IRTPRO does not calculate the MLE; however, the prior mean and variance for the MAP were set to 0 and 100, respectively, so that the resulting MAP estimates approximate the MLE.

Q1 is calculated as

$$Q_{1i} = \sum_{j=1}^{J} \frac{N_{ij}(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}$$

where N_{ij} is the number of examinees in cell j for item i, O_{ij} and E_{ij} are the observed and predicted proportions of examinees in cell j for item i. The expected or predicted proportion is calculated as

$$E_{ij} = \frac{1}{N_{ij}} \sum_{aej}^{N_{ij}} P_i(\hat{\theta}_a)$$

where $P_i(\hat{\theta}_a)$ is the item characteristic function for item i and examinee a. The summation is taken over examinees in cell j. The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$$gen\ Q_{1i} = \sum_{j=1}^{J} \sum_{k=1}^{m_i} \frac{N_{ij}(O_{ikj} - E_{ikj})^2}{E_{ikj}}$$

with

$$E_{ikj} = \frac{1}{N_{ij}} \sum_{aej}^{N_{ij}} P_{ik}(\hat{\theta}_a).$$

Both the Q1 and Generalized Q1 results are transformed into the statistic ZQ1, and are compared to a criterion, ZQ_{crit}, to determine acceptable fit.

$$ZQ_1 = \frac{Q_1 - df}{\sqrt{2df}}$$

and

$$ZQ_{crit} = \frac{N}{1500} * 4,$$

where Q is either Q1 or Generalized Q1 and *df* is the degrees of freedom for the statistic. The degrees of freedom are calculated as $J * (K - 1) - m$ where J is the trait interval, K is the number of score categories, and m is the number of estimated item parameters in the IRT model. In Yen (1981), the trait interval of 10 is used. For example, multiple choice items have $df = 10 * (2 - 1) - 3 = 7$. Poor fit is indicated where ZQ1 is greater than ZQ_{crit}.

The number of items flagged by Q1 can be found in Appendix A for operational items and Appendix B for field-test items.

Overall, few operational items were flagged by Q1. Algebra 1 had the most items flagged, with a total of six flags; however, these flagged items appeared across the four different core forms. Items flagged by Q1 were reviewed by psychometricians and content specialists before a final decision was made about their inclusion for student score calculation.

Appendix B lists the number of field-test items by grade and subject flagged by Q1. Before field-test items are placed onto forms for operational use in future administrations, they will be reviewed by content specialists and psychometricians. More information about test construction and item review can be found in Volume 2.

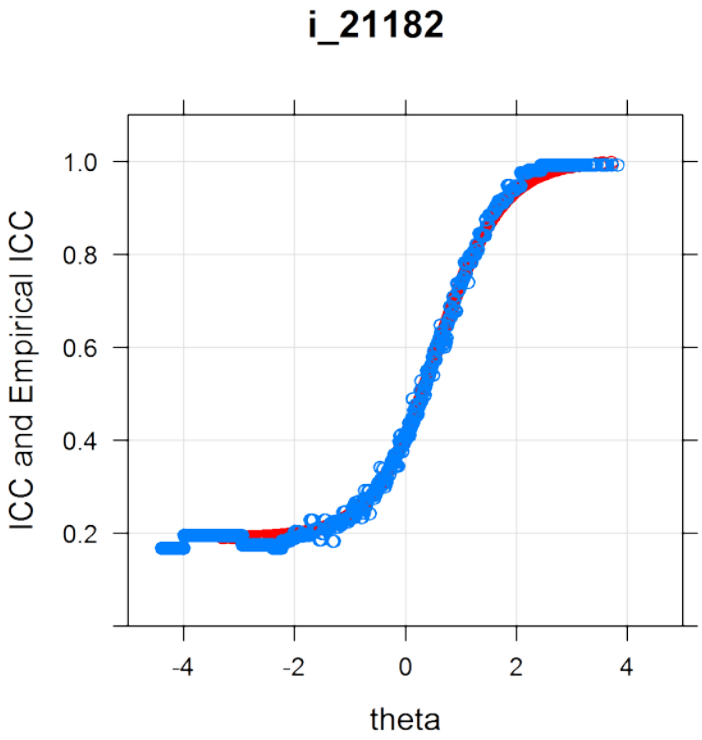
6.3.2 Item Fit Plots

Another way to evaluate item fit is to examine empirical fit plots for each item. The plots in this section are only examples of the types of fit plots used during item calibrations to add to the collection of evidence to evaluate item quality.

Fit plots were created for all items during calibration and are available upon request. Along with classical item statistics and Q1 flags, item fit plots were used to review items.

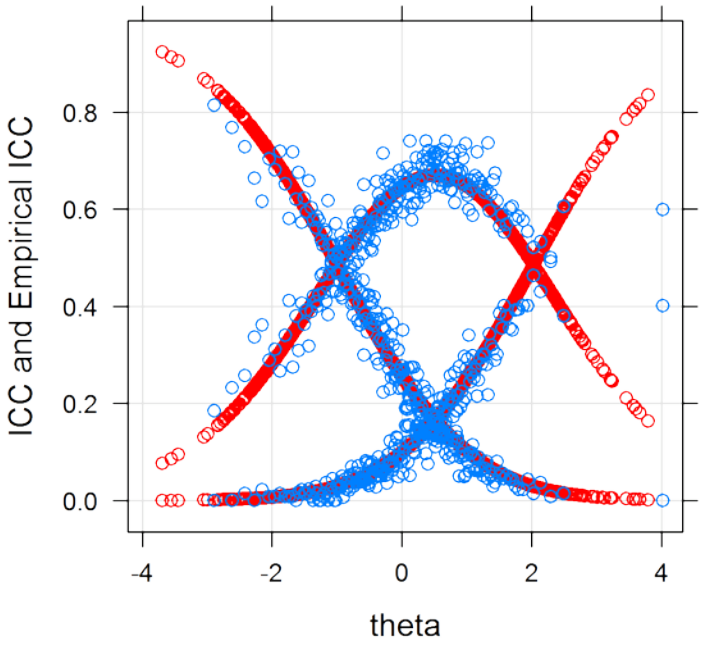
The fit plot in Figure 1 illustrates a one-point item that fits the item response model well. The dots represent the proportion of students within a score bin correctly answering the item. The solid line is the IRT-based item characteristic curve. A "good" item is one in which the dots are essentially superimposed over the line across the range of ability. In fact, the solid line is almost not visible underneath the dots for the first plot.

Figure 1: Example Fit Plot—Good Fitting One-Point Item



The plot in Figure 2 is provided for items worth two or more points. Again, the red lines are the IRT-based item characteristic curve. Here the dots represent the percentage of students, within a score bin, at each score point. Similar to the first plot, a "good" item is one in which the dots follow the solid lines across the range of ability.

Figure 2: Example Fit Plot—Good Fitting Two-Point Item



6.4 Results of Calibrations

This section presents a summary of the results from the classical item analysis and IRT analysis described in Section 5 for the 2017 spring operational and field-test items. The summaries here are aggregates; item-specific details are found in the appendices.

Tables 13–15 provide summaries of the *p*-values by percentile as well as the range by grade and subject for operational items. Note that the column *Total OP Items* shows the number of items that were used in the computation of the percentiles after excluding the dropped items. As noted in Section 1.4 above, there were multiple operational forms for EOC assessments. The summaries in Table 14 combine operational items across all forms. The field-test item summaries can be found in Appendix B; note that grade 3 Reading did not have any field-test items.

Table 13: Operational Item p*-Value Five-Point Summary and Range, Mathematics*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	54	0.14	0.30	0.61	0.74	0.83	0.88	0.93
4	54	0.21	0.27	0.57	0.71	0.81	0.92	0.94
5	54	0.22	0.30	0.42	0.55	0.68	0.80	0.84
6	55	0.10	0.19	0.34	0.52	0.68	0.84	0.87
7	56	0.11	0.17	0.31	0.40	0.57	0.76	0.87
8	56	0.12	0.13	0.20	0.30	0.59	0.77	0.83

Table 14: Operational Item p-*Value Five-Point Summary and Range, EOC*

Grade	Total OP Items*	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Algebra 1	126	0.02	0.05	0.16	0.33	0.50	0.69	0.78
Algebra 2	100	0.03	0.06	0.17	0.26	0.42	0.57	0.69
Geometry	120	0.03	0.06	0.13	0.24	0.43	0.67	0.78

*Note that operational items across all forms were combined.

Table 15: Operational Item p-*Value Five-Point Summary and Range, ELA*

Grade	Total OP Items	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	50	0.08	0.23	0.45	0.58	0.72	0.93	0.96
4	53	0.17	0.35	0.52	0.65	0.73	0.85	0.91
5	53	0.25	0.32	0.44	0.60	0.72	0.81	0.87
6	55	0.20	0.32	0.42	0.60	0.73	0.85	0.95
7	55	0.21	0.26	0.44	0.56	0.70	0.82	0.92
8	55	0.26	0.34	0.46	0.61	0.72	0.90	0.92
9	57	0.32	0.34	0.50	0.63	0.73	0.81	0.89
10	57	0.27	0.38	0.48	0.58	0.70	0.87	0.90

Tables 16–21 give the 3PL and 2PL item parameter summaries for Mathematics, EOC, and Reading. If fewer than 10 items existed in a model type for a given test, only the minimum and maximum are given. There were three or fewer GPCM items in any given grade and subject, and given the small number of items, summaries are not given.

Table 16: 3PL Operational Item Parameter Five-Point Summary and Range, Mathematics

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.58	0.61	0.76	0.92	1.17	1.46	1.77
	b	-2.33	-2.01	-1.48	-0.82	-0.28	0.16	0.66
	c	0.02	0.03	0.08	0.14	0.26	0.39	0.45
4	a	0.58	0.79	0.86	1.00	1.22	1.39	1.57
	b	-2.35	-1.97	-1.14	-0.49	-0.16	0.48	0.83
	c	0.05	0.07	0.13	0.19	0.26	0.38	0.41
5	a	0.64	0.69	1.04	1.15	1.19	1.28	1.31
	b	-1.09	-0.96	-0.34	0.08	0.35	0.87	1.30
	c	0.002	0.003	0.01	0.01	0.13	0.39	0.47

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
6	a	0.52	0.57	0.71	0.97	1.23	1.51	1.62
	b	-1.86	-1.83	-0.70	0.14	0.69	1.23	1.65
	c	0.001	0.003	0.01	0.05	0.18	0.28	0.33
7	a	0.36	0.51	0.83	1.03	1.19	1.42	1.63
	b	-2.23	-1.10	-0.14	0.36	0.75	1.56	1.69
	c	0.002	0.004	0.01	0.04	0.11	0.33	0.37
8	a	0.39	0.44	0.71	0.88	1.00	1.33	1.61
	b	-1.68	-1.24	-0.18	0.86	1.25	1.72	2.18
	c	0.001	0.001	0.01	0.08	0.19	0.25	0.45

Table 17: 2PL Operational Item Parameter Five-Point Summary and Range, Mathematics

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.48	0.73	0.95	1.08	1.20	1.39	1.44
	b	-1.48	-1.14	-0.85	-0.23	0.55	1.29	1.71
4	a	0.38	0.55	0.69	0.92	1.14	1.39	1.45
	b	-2.39	-1.48	-1.04	-0.24	0.27	1.18	1.26
5	a	0.59	0.73	0.81	0.89	1.04	1.26	1.40
	b	-1.47	-1.45	-0.67	-0.23	0.47	0.92	1.25
6	a	0.58	0.63	0.84	0.96	1.02	1.34	1.57
	b	-1.56	-1.41	-0.77	-0.09	0.56	1.00	1.22
7	a	0.54	0.60	0.78	1.08	1.36	1.38	1.40
	b	-0.46	-0.01	0.47	0.65	1.14	1.38	1.43
8	a	0.42	0.42	0.48	0.66	0.77	0.85	0.86
	b	-1.82	-1.57	0.54	1.43	1.88	2.23	2.63

Table 18: 3PL Operational Item Parameter and Five-Point Summary and Range, EOC

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Algebra 1	a	0.34	0.54	0.82	1.01	1.29	1.62	1.79
	b	-1.28	-1.05	0.16	0.90	1.24	1.93	2.02
	c	0.006	0.009	0.03	0.12	0.23	0.35	0.38
Algebra 2	a	0.44	0.63	0.96	1.15	1.42	1.66	1.94
	b	-0.22	0.14	0.83	1.11	1.58	1.87	2.13
	c	0.0003	0.001	0.01	0.11	0.21	0.34	0.37

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Geometry	a	0.35	0.57	0.85	1.11	1.41	1.86	1.96
	b	-1.17	-0.53	0.18	0.97	1.40	1.85	2.19
	c	0.004	0.006	0.01	0.07	0.18	0.33	0.44

Table 19: 2PL Operational Item Parameter Five-Point Summary and Range, EOC

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
Algebra 1	a	0.45	0.71	0.97	1.17	1.38	1.94	2.24
	b	-0.03	0.13	0.98	1.56	2.15	2.52	2.79
Algebra 2	a	0.65	0.77	0.94	1.20	1.47	1.84	1.99
	b	0.68	0.86	1.27	1.51	2.08	2.50	2.90
Geometry	a	0.49	0.64	0.89	1.18	1.51	1.73	1.85
	b	-0.70	0.45	1.03	1.33	1.73	2.34	3.48

Table 20: 3PL Operational Item Parameter Five-Point Summary and Range, ELA

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.38	0.51	0.75	0.97	1.12	1.35	1.84
	b	-2.24	-2.00	-0.79	-0.15	0.60	1.23	1.46
	c	0.01	0.04	0.10	0.16	0.22	0.29	0.42
4	a	0.30	0.42	0.58	0.70	0.96	1.21	1.37
	b	-1.73	-1.39	-0.91	-0.30	0.39	1.24	1.52
	c	0.01	0.01	0.05	0.11	0.20	0.31	0.47
5	a	0.37	0.48	0.63	0.76	0.91	1.11	1.25
	b	-2.14	-1.23	-0.53	-0.19	0.56	1.19	1.82
	c	0.01	0.01	0.05	0.18	0.23	0.43	0.50
6	a	0.45	0.47	0.57	0.70	0.93	1.19	1.38
	b	-3.11	-1.80	-0.79	-0.29	0.41	1.44	1.94
	c	0.01	0.01	0.05	0.13	0.28	0.36	0.51
7	a	0.41	0.45	0.59	0.70	0.92	1.17	1.30
	b	-2.59	-1.51	-0.52	0.15	0.56	1.31	1.53
	c	0.003	0.01	0.07	0.17	0.25	0.35	0.57
8	a	0.35	0.48	0.61	0.87	1.00	1.26	1.54
	b	-3.25	-2.01	-0.76	-0.34	0.39	1.15	1.33
	c	0.004	0.01	0.08	0.13	0.21	0.43	0.47

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
9	a	0.36	0.43	0.59	0.73	0.95	1.29	1.85
	b	-1.85	-1.28	-0.78	-0.09	0.49	1.14	1.33
	c	0.003	0.02	0.08	0.15	0.25	0.35	0.47
10	a	0.36	0.47	0.65	0.83	1.01	1.20	1.38
	b	-2.29	-1.36	-0.45	0.01	0.48	0.77	1.39
	c	0.01	0.02	0.08	0.24	0.30	0.38	0.49

Table 21: 2PL Operational Item Parameter Five-Point Summary and Range, ELA

Grade	Parameter	Min	5th Percentile	25th Percentile	50th Percentile	75th Percentile	95th Percentile	Max
3	a	0.43	0.43	0.45	0.55	0.73	0.95	0.95
	b	-0.11	-0.01	0.54	1.04	1.64	2.26	2.58
4* No Items	a							
	b							
5	a	0.84						1.09
	b	-1.54						-0.85
6	a	0.56						1.03
	b	-0.11						1.58
7	a	0.56						0.87
	b	-0.25						0.70
8	a	0.51						0.96
	b	-0.41						1.30
9	a	0.65						0.76
	b	-0.58						0.13
10	a	0.51						0.82
	b	-0.66						0.99

7. SUMMARY OF ADMINISTRATION

7.1 ITEM AND TEST CHARACTERISTIC CURVES

An item characteristic curve (ICC) shows the probability of a correct response as a function of ability given an item's parameters. Test characteristic curves (TCCs) can be constructed as the sum of ICCs for the items included on the test. The TCC can be used to determine examinee raw scores or percent-correct scores that are expected at given ability levels. When two tests are developed to measure the same ability, their scores can be equated through the use of TCCs. As such, it is useful to use TCCs during test construction. Items are selected for a new form so that the new form's TCC matches the target form's TCC as closely as possible.

The figures in Appendix D show the TCCs by grade and subject based on the final operational item parameters from the spring 2017 calibrations.

7.2 ESTIMATES OF CLASSIFICATION CONSISTENCY

See Classification Accuracy results in Section 3.4 of the 2016–2017 FSA Technical Report, Volume 4, Evidence of Reliability and Validity.

7.3 REPORTING SCALES

For spring 2017, the FSA ELA, Mathematics, and EOC tests report scale scores for each student. The score is based on the operational items presented to the student. Section 8.1 describes exactly how scores were computed.

Appendix E provides a summary of scale scores.

8. SCORING

8.1 FSA SCORING

8.1.1 Maximum Likelihood Estimation

The FSA tests were based on the 3-parameter logistic model (3PL) and Generalized Partial Credit Models (GPCM) of item response theory models, with the 2PL treated as a special case of the 3PL. Theta scores were generated using *pattern scoring*, a method that scores students differently depending on how they answer individual items.

Likelihood Function

The likelihood function for generating the maximum likelihood estimates (MLEs) is based on a mixture of items types and can therefore be expressed as

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR}$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N_{MC}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{exp \sum_{k=0}^{z_i} Da_i(\theta - \delta_{ki})}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h} Da_i(\theta - \delta_{ki})}$$

$$P_i = c_i + \frac{1 - c_i}{1 + exp[-Da_i(\theta - b_i)]}$$

$$Q_i = 1 - P_i$$

where c_i is the lower asymptote of the item response curve (i.e., the pseudo-guessing parameter), a_i is the slope of the item response curve (i.e., the discrimination parameter), b_i is the location parameter, z_i is the observed response to the item, *i* indexes item, *h* indexes step of the item, m_i is the maximum possible score point (starting from 0), δ_{ki} is the kth step for item *i* with m total categories, and $D = 1.7$.

A student's theta (i.e., MLE) is defined as $\arg\max_{\theta} log\big(L(\theta)\big)$ given the set of items administered to the student.

Derivatives

Finding the maximum of the likelihood requires an iterative method, such as Newton-Raphson iterations. The estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} \Big/ \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t}$$

where

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} = \frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} = \sum_{i=1}^{N_{3PL}} Da_i \frac{(P_i - c_i)Q_i}{1 - c_i}\left(\frac{z_i}{P_i} - \frac{1 - z_i}{Q_i}\right)$$

$$\frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} = -\sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1 - c_i)^2}\left(1 - \frac{z_i c_i}{P_i^2}\right)$$

$$\frac{\partial \ln L(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N_{CR}} Da_i \left(exp\left(\sum_{k=1}^{z_i} Da_i(\theta - \delta_{ki})\right)\right)\left(\frac{z_i}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} - \frac{\sum_{j=1}^{m_i} jexp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{\left(1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)\right)^2}\right)$$

$$\frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} jexp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}\right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}\right)$$

and where θ_t denotes the estimated θ at iteration *t*. N_{CR} is the number of items that are scored using the GPCM model and N_{3PL} is the number of items scored using 3PL or 2 PL model.

Standard Errors of Estimate

When the MLE is available, the standard error of the MLE is estimated by

$$se(\hat{\theta}) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta}\right)}}$$

where

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\left(\frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1 - c_i)^2} \left(1 - \frac{z_i c_i}{P_i^2}\right)$$

where N_{CR} is the number of items that are scored using the GPCM model and N_{3PL} is the number of items scored using 3PL or 2 PL model.

Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded and an MLE cannot be generated. In addition, when a student's raw score is lower than the expected raw score due to guessing, the likelihood is not identified. For FSA scoring, the extreme cases were handled as follows:

i. Assign the Lowest Obtainable Theta (LOT) value of –3 to a raw score of 0.
ii. Assign the Highest Obtainable Theta (HOT) value of 3 to a perfect score.
iii. Generate MLE for every other case and apply the following rule:
    a. If MLE is lower than –3, assign theta to –3
    b. If MLE is higher than 3, assign theta to 3

Standard Error of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the standard error (SE) is estimated based on Fisher information.

When the MLE is not available (such as for extreme score cases) or the MLE is censored to the LOT or HOT, the standard error (SE) for student *s* is estimated by

$$se(\theta_s) = \frac{1}{\sqrt{I(\theta_s)}}$$

where $I(\theta_s)$ is the test information for student *s*. The FSA tests included items that were scored using the 3PL, 2PL, and GPCM from IRT. The 2PL can be visualized as either a 3PL item with no pseudo-guessing parameter or a dichotomously scored GPCM item. The test information was calculated as

$$I(\theta_s) = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left(\frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} - \left(\frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{Q_i}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2 \right)$$

where, N_{CR} is the number of items that are scored using the GPCM model and N_{3PL} is the number of items scored using 3PL or 2 PL model.

For standard error of LOT/HOT scores, theta in the formula above is replaced with the LOT/HOT values.

8.1.2 Scale Scores

There are two scale types created for the FSA:

- A vertical scale score for ELA grades 3–10 ELA and grades 3–8 Mathematics
- A within-test scaled score for EOC tests

Table 22 shows the theta to scaled score transformation equations.

Table 22: Theta to Scale Score Transformation Equations

Subject	Grade	Theta to Scale Score Transformation
ELA	3	Scale Score= round(theta *20.000000 + 300.000000)
ELA	4	Scale Score = round(theta *20.237420 + 311.416960)
ELA	5	Scale Score = round(theta *21.230040 + 320.961420)
ELA	6	Scale Score = round(theta *21.861120 + 325.061500)
ELA	7	Scale Score = round(theta *21.581900 + 332.124320)
ELA	8	Scale Score = round(theta *21.531360 + 338.432720)
ELA	9	Scale Score = round(theta *21.751840 + 341.749740)
ELA	10	Scale Score = round(theta *21.284300 + 348.328540)
Mathematics	3	Scale Score= round(theta *20.000000 + 300.000000)
Mathematics	4	Scale Score = round(theta *20.899320 + 313.617800)
Mathematics	5	Scale Score = round(theta *22.050760 + 321.802560)
Mathematics	6	Scale Score = round(theta *21.684500+ 325.299220)
Mathematics	7	Scale Score = round(theta *20.379620 + 330.157540)
Mathematics	8	Scale Score = round(theta *19.952780 + 332.946420)
Algebra 1		Scale Score= round(theta *25.000000 + 500.000000)
Algebra 2		Scale Score= round(theta *25.000000 + 500.000000)
Geometry		Scale Score= round(theta *25.000000 + 500.000000)

When calculating the scale scores, the following rules were applied:

1. The same linear transformation was used for all students within a grade.
2. Scale scores were rounded to the nearest integer (e.g., 302.4 to 302; 302.5 to 303).
3. A standard error was provided for each score, using the same set of items used to derive the score.

The standard error of the scaled score is calculated as:

$$se(SS) = se(\theta) * slope$$

where $slope$ is the slope from the theta to scaled score transformation equation in Table 22.

8.1.3 Performance Levels

Each student is assigned a performance category according to his or her accountability scale score. Tables 23–25 provide the cut scores for performance standards for ELA, Mathematics, and EOC.

Table 23: Cut Scores for ELA by Grade

Grade	Cut between Levels 1 and 2	Cut between Levels 2 and 3	Cut between Levels 3 and 4	Cut between Levels 4 and 5
3	285	300	315	330
4	297	311	325	340
5	304	321	336	352
6	309	326	339	356
7	318	333	346	360
8	322	337	352	366
9	328	343	355	370
10	334	350	362	378

Table 24: Cut Scores for Mathematics by Grade

Grade	Cut between Levels 1 and 2	Cut between Levels 2 and 3	Cut between Levels 3 and 4	Cut between Levels 4 and 5
3	285	297	311	327
4	299	310	325	340
5	306	320	334	350
6	310	325	339	356
7	316	330	346	360
8	322	337	353	365

Table 25: Cut Scores for EOC

Grade	Cut between Levels 1 and 2	Cut between Levels 2 and 3	Cut between Levels 3 and 4	Cut between Levels 4 and 5
Algebra 1	487	497	518	532
Algebra 2	497	511	529	537
Geometry	486	499	521	533

8.1.4 Alternate Passing Score (APS)

The alternate passing score (APS) is the FCAT 2.0-equivalent score reported as an FSA scaled score. When Grade 10 ELA and EOC cut scores were reported in 2015, there was no approved FSA reporting scale, and so cut scores were reported as an FCAT 2.0 equivalent. The FSA scale transformation constants are now known, and so the passing scores can be reported on the FSA scale. Since the cuts recommended from the summer 2015 standard setting process have been approved, it is important to note that these APS cuts are used only with students who are retaking the test.

Equipercentile linking was used to find the FCAT 2.0 linked score, and this methodology relied on using an FCAT-looking score. The FCAT-looking score is the student's MLE transformed to be on a scale that uses the same transformation constants as the FCAT 2.0. Let $\hat{\theta}_s$ denote the FCAT-looking score for test *s* from the 2015 linking score conversion table. The APS is then found as

$$APS_{algebra} = \left[\frac{\hat{\theta}_{alg} - 400}{25}\right] * 25 + 500$$

$$APS_{geometry} = \left[\frac{\hat{\theta}_{geo} - 400}{25}\right] * 25 + 500$$

$$APS_{ela} = \left[\frac{\hat{\theta}_{e} - 244.870126}{18.822290}\right] * 21.284300 + 348.328540$$

The FSA score that corresponds to the cut score used for passing in 2015 is then found. These scores are below in Table 26.

Table 26: Alternate Passing Score Cut Points

Test	APS	FCAT-Linked Score	FCAT 2.0 and NGSSS EOC Looking Scales
Grade 10 ELA	349	245	245
Algebra 1	489	399	389
Geometry	492	396	392

Note that a student's passing indicator is based on whether or not the scale score meets the passing requirement, whereas the performance level is based exclusively on the scale score and the scale score cut point.

In Grade 10 ELA, the APS is 349 and the scale score cut point for Level 3 is 350. If a Grade 10 ELA student scores 349, he or she receives a passing status of Y and a performance level of 2.

More information can be found in Section 6.3 of Volume 1 of the 2014–2015 Technical Reports.

8.1.5 Reporting Category Scores

In addition to overall scores, students also receive scores on reporting categories. Let b_{sq} represent the subset of operational items presented to student s in reporting category q. Students will receive a raw score for each reporting category, with these scores being derived using only b_{sq}. That is, the raw score is calculated as the sum of the scores on the subset of operational items measuring reporting category q. The number of raw score points for each test and reporting category is provided in Appendix F, along with summaries of scores from spring 2017.

9. STATISTICAL SUMMARY OF TEST ADMINISTRATION

9.1 DEMOGRAPHICS OF TESTED POPULATION, BY ADMINISTRATION

Tables 27–29 present the distribution of students, in counts and in percentages, who participated in the spring administration of the 2016–2017 FSA by grade and subject. The numbers presented here are based on the reported status in the approved spring SSR files. The subgroups reported here are gender, ethnicity, students with disabilities (SWD), and English language learners (ELL).

Table 27: Distribution of Demographic Characteristics of Tested Population, Mathematics

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
3	*N*	228745	110867	117878	52024	77998	83559	24679	34970
	%	100	48.47	51.53	22.74	34.10	36.53	10.79	15.29
4	*N*	210120	103797	106323	44504	70827	80545	22087	19220
	%	100	49.40	50.60	21.18	33.71	38.33	10.51	9.15
5	*N*	214004	104886	109118	46326	72281	81324	25888	20305
	%	100	49.01	50.99	21.65	33.78	38.00	12.10	9.49
6	*N*	196833	96946	99887	43105	65546	75692	23352	14421
	%	100	49.25	50.75	21.90	33.30	38.45	11.86	7.33
7	*N*	179064	87452	91612	39496	59453	68855	20954	12652
	%	100	48.84	51.16	22.06	33.20	38.45	11.70	7.07
8	*N*	132952	63430	69522	33324	46119	46557	19466	11769
	%	100	47.71	52.29	25.06	34.69	35.02	14.64	8.85

Table 28: Distribution of Demographic Characteristics of Tested Population, EOC

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
Algebra 1	*N*	208197	102196	106001	41918	68909	83356	17698	12430
	%	100	49.09	50.91	20.13	33.1	40.04	8.50	5.97
Geometry	*N*	180301	90929	89372	36385	58438	73161	14688	8924
	%	100	50.43	49.57	20.18	32.41	40.58	8.15	4.95
Algebra 2	*N*	122638	65443	57195	21445	36430	55119	5062	2653
	%	100	53.36	46.64	17.49	29.71	44.94	4.13	2.16

Table 29: Distribution of Demographic Characteristics of Tested Population, ELA

Grade	Group	All Students	Female	Male	African-American	Hispanic	White	SWD	ELL
3	*N*	228166	110642	117524	51926	77621	83499	24548	34346
	%	100	48.49	51.51	22.76	34.02	36.6	10.76	15.05
4	*N*	207703	102705	104998	43936	69793	79869	21834	18342
	%	100	49.45	50.55	21.15	33.6	38.45	10.51	8.83
5	*N*	211546	103785	107761	45773	71148	80676	25580	19441
	%	100	49.06	50.94	21.64	33.63	38.14	12.09	9.19
6	*N*	200977	98989	101988	43116	66189	78191	23032	13776
	%	100	49.25	50.75	21.45	32.93	38.91	11.46	6.85
7	*N*	198891	98033	100858	41721	65253	78505	20901	12203
	%	100	49.29	50.71	20.98	32.81	39.47	10.51	6.14
8	*N*	198804	97475	101329	42174	65320	77963	20543	12113
	%	100	49.03	50.97	21.21	32.86	39.22	10.33	6.09
9	*N*	199934	98776	101158	42739	64458	79735	18789	11328
	%	100	49.40	50.6	21.38	32.24	39.88	9.40	5.67
10	*N*	198691	99494	99197	42562	63834	79524	17944	9847
	%	100	50.07	49.93	21.42	32.13	40.02	9.03	4.96

10. QUALITY CONTROL FOR DATA, ANALYSES, SCORING, AND SCORE REPORTS

10.1 DATA PREPARATION AND QUALITY CHECK

AIR's quality assurance procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at AIR.

Prior to any analysis, data were first extracted from the database of record (DoR). Processing and exclusion rules were then applied to determine the final data file to be used in psychometric analyses.

Once the data file was finalized, it was passed to two psychometricians who used the files for all analyses independently. Each psychometrician independently implemented the classical and IRT analyses. The results from the two psychometricians (e.g., the IRTPRO output files) were formally compared. Any discrepancies were identified and resolved.

When all classical and IRT results matched from the independent analysts, the results were uploaded to the secure file transfer protocol (SFTP) for review. FDOE psychometricians, HumRRO, and Buros also completed independent replications. During calibrations, daily calls were held with AIR, FDOE, TDC, HumRRO, and Buros to discuss classical statistics and IRT analyses. Content experts from AIR and TDC also reviewed classical statistics and gave input to the discussion. Results were approved by FDOE only when there was replication and verification from all parties.

The daily calibration calls were an important source for quality control and typically proceeded in an iterative fashion. Typically, two to three tests were evaluated during the calls, reviewing all of the evidence on item quality including classical analyses, IRT-based statistics and fit statistics, fit plots, and in many cases, reviewing the content of the item in a web-based setting.

During these calls, the teams discussed any observed issues or concerns with flagged items and determined if the item suffered from any content or statistical issues that warranted removing it from the set of core items used for scoring.

AIR uploaded item statistics to the item bank only after receiving final confirmation from all parties that the IRT statistics were accurate and that the items were appropriate for use in operational scoring.

10.2 SCORING QUALITY CHECK

Prior to the operational testing window, AIR's scoring engine was tested to ensure that the MLEs produced by the engine were accurate. This is a process referred to as *mock data*. During mock data, AIR established all systems and simulated item response data as if real students responded to the test items. AIR then tested all programs and verified all results before implementing the operational test. Simulated data was posted to the SFTP for FDOE, HumRRO, and Buros to allow all parties to test their systems.

Once final operational item calibrations were complete and approved by FDOE, item parameters were uploaded to AIR's item tracking system (ITS), and student scores—including MLEs, scale scores, and reporting category raw scores—were generated via the scoring engine.

Similar to the verification process with calibrations, independent score checks were performed by AIR, FDOE, and HumRRO. Scores were only approved by FDOE when there was a three-way replication and verification.

10.3 SCORE REPORT QUALITY CHECK

Two types of score reports were produced for the 2016–2017 FSA: online reports and printed reports. The FSA reporting system (FSA-R) provided the information on student performance and the aggregated summary at various levels (e.g., the district). The paper individual student reports (family reports) were provided to families of students who took the FSA tests.

Before deploying the 2016–2017 reports in FSA-R, various test cases were produced. The test cases were generated based on users' roles, functionality, and jurisdiction in the FSA-R. Each test case described a scenario and the expected result of the scenario. After all of the applicable test cases were executed successfully on the trial site without any issues, the codes were then deployed to the live site.

AIR also implemented a series of quality control steps to ensure error-free production of educator reports deployed via FSA-R and the paper family-score reports. To begin, using several types of dummy data, members from the AIR score reporting team compared proofs with mock-ups and communicated with the programmers to ensure that the reports were being generated as they should appear. These dummy data were created to test the accurate placement of all variables on the score reports and to review graphic alignment. After thoroughly testing the code using the dummy data, AIR then reviewed thousands of reports with live data to ensure full accuracy of the data.

The last quality assurance phase for the individual paper family score reports occurred at the print site. AIR provided training to print vendors on processes and procedures to ensure that the correct numbers of reports were printed, packaged, and shipped. Several AIR staff members also checked the reports as they were printed and packaged to ensure that they looked as they should and were packaged and shipped to the correct locations.

11. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Cohen, J., Johnson, E., & Angeles, J. (2000). *Variance estimation when sampling in two dimensions via the jackknife with application to the National Assessment of Educational Progress.* Washington DC: American Institutes for Research.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. *(2nd ed.)* New York, NY: Springer.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Michaelides, M.P., & Haertel, E.H. (2004). *Sampling of Common Items: An Unrecognized Source of Error in Test Equating* (CSE Report 636). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*(2)*,* 159–176.

Phillips, G.W. (2010). *Score drift: Why district and state achievement results unexpectedly bounce up and down from year to year.* Training session at the National Council for Measurement in Education, Denver, CO.

Somes, G. W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

van der Linden, W. J. and Hambleton, R. K. (Eds.) (1997) *Handbook of modern item response theory*. New York: Springer-Verlag.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement (*ETS Research Report No. 12-08). Princeton, NJ: Educational Testing Service.